

# VU Research Portal

## Refining Statistical Data on the Web

Merono Penuela, Albert

2016

### **document version**

Publisher's PDF, also known as Version of record

[Link to publication in VU Research Portal](#)

### **citation for published version (APA)**

Merono Penuela, A. (2016). *Refining Statistical Data on the Web*. [PhD-Thesis - Research and graduation internal, Vrije Universiteit Amsterdam].

### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

### **Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

### **E-mail address:**

[vuresearchportal.ub@vu.nl](mailto:vuresearchportal.ub@vu.nl)

## CONTENTS

---

1	INTRODUCTION	17
1.1	Historical Statistics . . . . .	20
1.2	Integration of Messy Spreadsheet Collections . . . . .	22
1.3	Data Quality and Transformation . . . . .	24
i	HISTORICAL STATISTICS	27
2	WHAT IS HISTORICAL DATA?	29
2.1	The Semantic Web . . . . .	30
2.2	Historical Research . . . . .	31
2.2.1	The Life Cycle . . . . .	32
2.2.2	Knowledge Discovery in Social History . . . . .	35
2.3	Historical Data . . . . .	38
2.3.1	A Classification of Historical Data . . . . .	38
2.3.2	An Ontological Framework . . . . .	43
2.4	Conclusion . . . . .	48
3	INTEGRATION PROBLEMS IN HISTORICAL STATISTICAL DATA	49
3.1	Integration Problems . . . . .	50
3.1.1	Integration Problems in Social History . . . . .	50
3.1.2	Integration Problems of Spreadsheets . . . . .	55
3.1.3	Integration Problems in History . . . . .	58
3.2	Related Work . . . . .	60
3.2.1	Provenance . . . . .	60
3.2.2	Data models . . . . .	61
3.2.3	Schema integration . . . . .	64
3.2.4	Data quality . . . . .	67
3.3	Conclusion . . . . .	69
ii	INTEGRATION OF MESSY SPREADSHEET COLLECTIONS	71
4	WEB-BASED INTEGRATION OF MESSY SPREADSHEET COLLECTIONS	73
4.1	Introduction . . . . .	74
4.2	Messy Spreadsheet Collections . . . . .	76
4.3	Integration of MSC on the Web . . . . .	77
4.3.1	Step 1: Data Location Definition . . . . .	78
4.3.2	Step 2: Dimension Conciliation . . . . .	80
4.3.3	Step 3: Measurement Transformation . . . . .	84

4.3.4	Step 4: Error Detection . . . . .	85
4.3.5	The Integrator . . . . .	86
4.4	Evaluation . . . . .	88
4.4.1	Use Case 1: the Dutch Historical Censuses . . . . .	89
4.4.2	Use Case 2: Wages, Prices and Welfare . . . . .	90
4.4.3	Use Case 3: UK Messy Open Data . . . . .	91
4.5	Related Work . . . . .	91
4.6	Discussion . . . . .	92
4.7	Conclusions and Further Work . . . . .	93
5	5-STAR LINKED HISTORICAL DUTCH CENSUS DATA . . . . .	95
5.1	Introduction . . . . .	96
5.2	The CEDAR project . . . . .	98
5.3	The Dutch Historical Censuses Dataset . . . . .	101
5.3.1	Previous Efforts . . . . .	102
5.3.2	Towards Linked Historical Dutch Census Data . . . . .	106
5.4	Data Conversion and Modelling . . . . .	108
5.4.1	Data Conversion . . . . .	108
5.4.2	Raw Data . . . . .	109
5.4.3	Integration Rules as Open Annotations . . . . .	111
5.4.4	Harmonized RDF Data Cube . . . . .	112
5.4.5	Provenance . . . . .	112
5.4.6	Named Graphs and URI Policy . . . . .	113
5.5	Linked Dataset Description . . . . .	114
5.5.1	Internal Links . . . . .	115
5.5.2	External Links . . . . .	116
5.6	Usage . . . . .	119
5.7	Impact and Availability . . . . .	124
5.7.1	Impact . . . . .	124
5.7.2	Availability . . . . .	127
5.8	Discussion . . . . .	128
iii	DATA QUALITY AND TRANSFORMATION . . . . .	131
6	QUALITY OF EVOLUTION IN DIACHRONIC WEB SCHEMAS . . . . .	133
6.1	Introduction . . . . .	134
6.2	Related Work . . . . .	135
6.3	Change Models for Diachronic Web Schemas . . . . .	136
6.3.1	Change Heuristic . . . . .	137
6.3.2	Feature Set . . . . .	137

6.3.3	Pipeline . . . . .	138
6.3.4	Quality of Evolution Metric . . . . .	139
6.4	Measuring Quality of Evolution . . . . .	140
6.4.1	Input Data . . . . .	140
6.4.2	Experimental Setup . . . . .	141
6.4.3	Results . . . . .	142
6.4.4	Characterization of Quality Version Chains . . . . .	143
6.5	Lessons Learned . . . . .	145
6.6	Future Work . . . . .	148
7	QUALITY OF WEB DATA CUBES: LINKED EDIT RULES . . . . .	149
7.1	Introduction . . . . .	150
7.2	Background and Problem Definition . . . . .	151
7.3	Related Work . . . . .	154
7.4	Approach . . . . .	155
7.4.1	Linked Edit Rules and RDF Data Cube . . . . .	155
7.4.2	From edit rules to Linked Edit Rules . . . . .	157
7.4.3	LER Architecture . . . . .	158
7.5	Implementation . . . . .	159
7.5.1	Stardog Linked Micro-Edit Rules . . . . .	160
7.5.2	Stardog Linked Macro-Edit Rules . . . . .	160
7.5.3	Stardog as Validation Proxy . . . . .	163
7.6	Evaluation . . . . .	164
7.7	Discussion and Future Work . . . . .	166
8	SCRY: EXTENDING SPARQL USING FEDERATION . . . . .	169
8.1	Introduction . . . . .	170
8.2	Problem Definition . . . . .	171
8.3	Related Work . . . . .	173
8.4	SCRY . . . . .	174
8.4.1	Typical Use . . . . .	174
8.4.2	Implementation . . . . .	178
8.4.3	Syntax . . . . .	179
8.4.4	Limitations . . . . .	181
8.5	Use Cases . . . . .	182
8.5.1	Statistics . . . . .	182
8.5.2	Bioinformatics . . . . .	184
8.6	Conclusions . . . . .	186
9	CONCLUSION . . . . .	189
9.1	Results . . . . .	189

9.1.1	Data Integration Problems in History . . . . .	189
9.1.2	Integration of Messy Spreadsheet Collections . . . . .	192
9.1.3	Data Quality and Transformation . . . . .	196
9.1.4	Answer to Main Research Question . . . . .	198
9.2	Limitations . . . . .	199
9.3	Lessons Learned and Future Work . . . . .	201

BIBLIOGRAPHY		207
--------------	--	-----